# Project Abstract

**Austronesian Societies: Reading Social Structure from the Genome**
0432262
J. Stephen Lansing and Joseph Watkins
University of Arizona

Michael Hammer and Tatiana Karafet
Genomics and Technology Core, Arizona Research Labs

## Overall Mission/Objective

Genetic and language modeling, data collection, and analysis can form an essential partnership with the traditional field work of cultural anthropology. Such a combination of approaches can shed important insights into the often nonlinear relationships among the various human social behavioral factors that can influence patterns of relatedness, such as kinship and marriage rules, migration, language drift and historical economic adaptations.

While most current research focuses on genetic structure at the population level, this project will use non-coding neutral genetic markers to study the emergence of patterns of relatedness among individuals within communities. Results from our previous research show that a change in mode of production (from tribal horticulture to irrigated rice production) leaves clear traces in the genetic patterns of small Balinese communities. This project will look for comparable genetic signatures in neighboring populations. The aim is to gain a better understanding of the major variables that affect patterns of relatedness among human groups.

To address these questions, this project, in cooperation with the Eijkman Institute and the National Language Institute in Jakarta, has been collecting DNA and language samples from communities throughout Indonesia where good archaeological data on the age and relatedness of settlements can be used to test assumptions about migrations and population histories based on genetics. In addition, the National Language Institute has provided 1000 word lists from dozens of villages. The Eijkman Institute is interested in medical genetics issues; this project is focused on "community" genetics. Thus, our goals are quite complementary.

New mathematics models and statistical analyses of the genetic data supplemented with computer simulations, e.g. agent-based modeling, are needed because standard population genetics models assume that the dynamics have reached equilibrium, an unwarrantable assumption for the short time spans of concern here.

## Progress and (Preliminary) Outcomes

The period from June to November, 2005 is primarily devoted to the collection of data – both genetic and linguistic – from a variety of parts of Indonesia. Here is report from the first two excursions.

**Sumba**: A total of about 420 blood samples were taken - approximately 50 samples from men in 8 Sumbanese villages. The villages chosen were based on two criteria: they represent different language areas, and they are considered old source villages, from which other villages later were created as daughter settlements. They are patrilocal, organized into clans, which practice preferential marriage to mother's brother's daughter. The more traditional the village, the more this preference is enforced—in extreme cases infant girls are already promised to their husbands. Local health workers who speak the local language(s) helped each subject fill in a questionnaire and consent form. We tried for 4 generations of pedigree information, or as much as they could remember, noting the origin village for wives, and how many kilometers away from the husband's origin village. Typically 3 to 7 clans occupy one of these old source villages, so it was not too hard to find men who are not brothers or first patricousins.

Inaccuracies were found in the Swadesh word lists for Sumba. Thus, native speakers from each language or dialect were videotaped pronouncing the 200 words on the list.

There was interesting variation in the social structure of the villages we studied – in Eastern Sumba, society is more hierarchical. Heads of clans have many wives (up to a dozen). They also own slaves. It will be interesting to see if these differences in contemporary social structure will show up as historical trends in the DNA data.

Finally, the blood samples taken in Sumba will be used for a hepatitis study, and the results reported back to the people via the public health offices.

**Flores.** Indonesian archaeologists led by Teuku Jacob report the existence of "pygmies", i.e. modern Austronesian-speaking farmers of very short stature, in the village of Rampasasa near the cave where a team of Australian and Indonesian archaeologists discovered the bones of "*Homo Floresiensis*" in 2003. These pygmies are very small modern Indonesians, culturally Manggarai.

We obtained blood samples from men in Rampasasa, including a handful of men short enough to qualify as pygmies. We also measured their height and weight. We also spent 9 days with people from the local public health department, collecting samples from pygmies in other villages. Interestingly, the pygmies are firmly convinced that they are the descendants of a small furry forest hunter who married a human woman. The pygmies are both men and women, and they often (but not always) marry one another. They look quite distinctive and do not seem to be just a collection of short farmers. The Eijkman Institute plans to look for abnormalities in thyroid function from the blood samples.

While we were collecting these samples, two teams were also trying to extract DNA from the bones discovered in the cave. One team is based at the Eijkman Institute; the other at the Max Planck Institute in Leipzig.

Throughout our travels in Flores, we asked people about the existence of very short people, but found no other pygmies until we reached a village in the Ende region of Central Flores, where

we found two possible pygmies (they are a bit taller) and obtained their blood, in a village where we also took 45 more samples from the men.

After the pygmies, probably the next most interesting samples came from the region of Bajawa, in the mountains of Central Flores. Two villages that were sampled trace their descent matrilineally, in contrast to their immediate neighbors who are typical patrilineal eastern Indonesians. In one matrilineal village, we took samples from women as well as men.

We have more to do on Flores including samples from the extreme eastern end of the island, and the adjoining islands in the direction of Alor, which is populated by people who speak Papuan languages (unrelated to the Austronesian languages of Flores).

In Tucson, we are working along three strands in preparation for the data.

**Linguistic Component.** The Indonesian lexical database, which has already been largely completed, but will be finalized with the entry of the new data currently being collected, the organization of lexical data and inferences based thereupon will proceed in three steps:

**Grouping of lexical items into cognate sets**. For example, the words for *fire* will be organized into groups, the words for *water* into groups, on the basis of superficial similarity. Typically, words which group together for one item will generally group for others, although this will not always be the case due to the complex nature of language change and the idiosyncracies of lexical replacement. We plan to develop a program which that search a lexical items and group them according to phonological similarity. The comparison can be done first at the segmental level, and then more finely at the featural level.

**Establishment of regular sound correspondences and reconstruction of the paths of sound change**. These correspondences will ultimately reconstruct the original phonetic values of the words in question (we are fortunate to have as a check the reconstructed proto-language ancestral to all Austronesian languages outside of Taiwan, Proto-Malayo-Polynesian). When this is completed, then it will be assessed how many languages show the same sound changes taken together as a group.

**Identification of loans and interpretation of sound change strata**. Loan words often stand out because they superficially appear to be inherited, and will fail to display the sound correspondences expected from the normal evolution of the language. The best explanation is that the word has been borrowed from a contiguous language. The second way in which contact can be discerned is through shared sound changes. If such changes between can be shown to be recent, but older sound changes are not shared, then the best explanation is that they share these more recent sound changes because they have come into contact with each other and the changes are part of an areal phenomenon.

**Genetics.** The specific aim is to develop a comparative genetic framework for different compartments in the human genome, Y-chromosome binary and microsatellite markers, mtDNA and autosomal short tandem repeats (STRs).

We have examined genetic variation on the non-recombining portion of the Y chromosome (NRY) to investigate the paternal population structure of hunter-gatherer population from Borneo, several isolated groups from Mentawai and two subaks from Bali. A set of 71 biallelic markers and 12 STRs on the NRY were genotyped.

We have developed two nanoplex PCR reactions to survey 18 autosomal microsatellite polymorphisms. The systems were chosen to insure that none of them are closely linked, maximizing statistical independence. While the bulk of progress towards autosomal STR typing is anticipated to be in year 2, we have typed some 200 samples from different Bali populations.

**Mathematical Modeling.** Recent results in both the biophysical and biochemical literature give very refined answers on the nature of microsatellite evolution. The biochemical experiment is based on direct observation of mutations in parent-child pairs. The biophysical assays are *in vitro* experiments to analyze DNA polymerase errors in copying microsatellite sequences. The suite of models suggested by these experiments are amenable, using the theory of Markov processes, to exact analytical expressions for the probability distributions for the change in microsatellite length over any given number of generations. This gives, in a precise statistical sense, best possible estimates for time to most recent common ancestor. A preprint is nearly completed with software implementation soon to follow. The next modeling project involves examining the social structures as a random network.

**Broader Impacts**

The project will result in the creation of new analytical tools, both statistical and computational, to investigate genetic population structure at the fine-grained level of resolution provided by microsatellites and will also provide a computational framework for integrating genetic analysis with models of language evolution. These tools will open new possibilities for the empirical study of fundamental anthropological questions well beyond the study region for this project.

More broadly, this project is quintessentially about marginalized populations. Only under the backdrop of solidly understood population genetics can the impact of selection and survivability be measured using medical genetics. In addition, a more foundational understanding of the kinship networks that are created to support economies can prove to be a cautionary tale to those transnational entities that seek to impose broad and blunt instruments for development. Thus, dissemination of these ideas and methods are central to the project. We have been invited to work with indigenous groups in the Philippines to extend these ideas to that region. In addition, a video about the project will be produced by a leading British film-maker.